PART 4

LM2-TOXIC

Appendix 4.4.1 Sample Data Interpolation for the LMMBP

Xiangsheng Xia
Computer Sciences Corporation
Large Lakes Research Station
9311 Groh Road
Grosse Ile, Michigan 48138

Many sample data sets of physical and chemical parameters collected for the Lake Michigan Mass Balance Project (LMMBP) were often sparse and occurred on irregular grids. For modeling purposes, values of these parameters were needed on a 5 x 5 km grid. This presented a problem of using sample data to estimate or predict values in areas which were not sampled. Thus, some interpolation mechanisms based on "insufficient" samples were needed to bridge the gap between the desired and the reality world of data collection. Distance square inverse and natural-neighbor interpolation methods were carefully studied and applied to sample data analysis for this project.

The distance-weighted-averaging approach determines the estimated values at grid points as the sum of weighted values of the individual sample datum. In general, the closer a datum point to the grid point to be estimated, the greater influence the datum at that point exerts. It is a method characterized as a global approach. The distance-weighted-averaging method is well understood and widely accepted by scientists in various fields. It is also easy to implement. The major disadvantage of this method has been its tendency to smooth out all small variations in the relatively small local area. Therefore, it is not very well suited to find the trend of samples in small local areas. The distance-weighted-averaging interpolation is compromised by its essentially one-dimensional nature. Although the interpolating surface is smooth, it cannot, for instance, form ridges or domes from sparse data. Furthermore, distance-weighted-averaging is unable to infer (or extrapolate) a surface lying above or below the range of sample values. In general, the estimation computed by distance-weighted-averaging lies between the maximum and minimum of the sample data.

Neighborhood-based interpolation, on the other hand, is a local approach which utilizes all the (natural) neighbors of the sample points. The natural-neighbor method can infer values outside the known range. It is unique for a given data configuration and choice of blending function parameters. If used properly on dense data sets, neighborhood-based interpolation can provide much richer information such as rapid changes, ridges, or dams in smaller areas. However, neighborhood-based interpolation, in contrast to distance-based methods, is much more complicated to implement and harder to understand. In case an ambiguity or unexpected phenomena arise from a neighborhood-based interpolation, it may require a knowledgeable person to make reasonable interpretation of results.

During the course of the LMMBP data analysis process, distance square inverse interpolation combined with application codes written in Interactive Data Language (IDL) were used intensively to find the interpolated values of a 5 x 5 km grid of Lake Michigan for various parameters,

such as polychlorinated biphenyls (PCBs), atrazine, nutrients, etc. On the other hand, natural-neighbor interpolation was primarily used for sediment data analysis where sample locations were relatively dense.

More details of distance square inverse and natural-neighbor interpolations are presented in the next two sections. Many applications of interpolation have been developed, including contour plots, volume-weighted averages, and others. These are discussed in Section 4.4.1.4. Some problems applying natural-neighbor interpolation are discussed in Section 4.4.1.5.

4.4.1.1 The Distance Square Inverse Method

The inverse distance to a power method is a weighted-average interpolation. Data are weighted during interpolation such that the influence of one sample point relative to another declines with increasing distance from the grid node. Weighting is assigned to data using a weighting power that controls how the weighting factor drops off as distance from a grid node increases. As the power increases, the grid node value approaches the value of the neighboring data points. The weighting power determines how quickly weighting falls off with distance from the grid node. As the power approaches zero, the generated surface approaches a horizontal planar surface through the average of all observations from the data file. As the power increases, the generated surface is a "nearest neighbor" interpolation, and the resultant surface becomes polygons which represent the nearest observation to the interpolated node. Power values are usually between one and three to avoid extreme results. Distance square inverse is the distance-weighted method with the power chosen as two.

The smoothing factor parameter allows one to incorporate an uncertainty factor associated with sample data. The larger the smoothing factor parameter, the less influence a particular observation has in computing a neighboring grid. The smoothing factor for this study was 2.5 (miles).

The equation in the inverse distance square method is:

$$v_i = \frac{\sum_{j=1}^{n} \left(\frac{1}{d_{i,j}^2 + r_o^2} \cdot C_j \right)}{\sum_{j=1}^{n} \left(\frac{1}{d_{i,j}^2 + r_o^2} \right)} \qquad \text{(A4.4.1.1)}$$

where

j = runs for all samples

v_i = interpolated value at grid point i,

C_j = value of sample j,

$d_{i,j}$ = distance between grid point i and sample location j,

r_o = smoothing factor

n = total number of samples being considered in the interpolation

An IDL code which implements the inverse distance square interpolation scheme was received from David Schwab (National Oceanic and Atmospheric Administration, Great Lakes Environmental Research Laboratory). This was further developed for the LMMBP data analysis.

4.4.1.2 The Natural-Neighborhood Method

Natural-neighbor interpolation offers a different approach to spatial interpolation and extrapolation. It has good mathematical properties and offers more flexibility than the distance square inverse method.

All interpolation methods involve, to some extent, the idea that the value of the interpolated point should depend more on data values at nearby data sites than at distant ones. In natural-neighbor interpolation, the idea of neighbors in a spatial configuration is formalized in a natural way and made quantitative, and the properties of the method depend on an apparently new geometrical identity relating this quantitative measure of neighbors to position.

Any two data are natural-neighbors if there is a location or region that is equally close to each of the pair, and no other datum is closer. Any three or more data on the plane are natural-neighbors if no other datum lies within their circum-circle. The spatial relationships determined by a set of natural-neighbors circles have two common and widely known graphical representations. These are Voronoi tessellation (of Voronoi polygons) and Delaunay triangulations. The Voronoi tessellation illustrates that each datum has a unique natural-neighbor region associated with it and is bounded by halfway interfaces of that datum with its natural-neighbor. Neighborhood coordinates are local coordinates relating the position of the interpolation point to reach a datum in the neighborhood subset. These coordinates (weights for the interpolation), ranging between zero and one, are proportional to areas defined on natural-neighbor regions for each of the data. Such coordinates are superior to distance-based coordinates. Distance-based coordinates make no allowance for the distances to the other data; that is, distance-based interpolation is not sensitive to a changing spatial context. Finally, natural-neighbor interpolation is a linear-weighted average of natural-neighbor coordinates. The basic equation used in natural-neighbor interpolation can be defined as follows:

$$V_i = \sum_{j=1}^{k} \left(W_j \cdot S_j \right) \qquad \text{(A4.4.1.2)}$$

where

V_i = interpolated value at grid point i,

k = number of samples inside the natural-neighborhood of V_i,

Sj = value of sample j,

W_j = weight associated with S_j.

The c-code nngridr, a complete commercial package of the natural-neighbor algorithm, from David Watson (Watson, 1994), with some modification by in-house developers, was used for developing applications of natural-neighbor interpolation at the Large Lakes Research Station (LLRS).

4.4.1.3 Application

Interpolations, distance square inverse or natural-neighbor, were needed to build two-dimensional estimates of a 5 x 5 km grid of Lake Michigan from a limited number of samples of various parameters. Other applications could then be applied based on the interpolated grid data.

4.4.1.3.1 Contouring Plots

Interpolated grid data is a list of numbers representing the estimates of physical parameters on the grid for each grid point. Contour plots connect points in the grid having the same value with lines. An incredible amount of information about the data can be revealed by contour plots. These include plateaus and canyons, trends, the existence and location of high and low concentrations, etc.

Contour plots are very effective visualization tools for analyzing data. Contours in this study were created by using IDL and other tools. There were a surprising variety of approaches used to generate contours. The various techniques that were applied possess their own advantages and disadvantages. IDL's standard CONTOUR procedure uses grid contouring which is the most widely used contouring technique (Research Systems, Inc., 1995, see Section 15-1). CONTOUR generates plots from data stored in a rectangular array (grid data) which usually is generated by interpolation and extrapolation. Some other information such as the boundary of Lake Michigan, sample locations, and the maximum and minimum values for samples were also produced.

4.4.1.3.2 Volume-Weighted Averaging With Formulations

One way to evaluate and validate the performance of mathematical models is to compare the model output and the measured data at the same time (cruises) and same location (segments). Volume-weighted average (VWA) is a method to compute the estimated field data associated with a segment and a cruise. Depending on the model and segmentation scheme used, a segment consists of cells of 5 x 5 km at certain depth range called a layer. The locations of cells associated with segments are normally

provided by segmentation files. The volume concentration for one cell can be computed by multiplying interpolated concentration of this cell by its volume. The volume concentration for a segment is the sum of volume concentration of all cells in this segment. And finally, VWA can be computed by dividing the volume concentration of the segment by the total segment volume. The equation for computing VWA is :

$$A_n = \frac{\sum_{i=1}^{n} \left(V_i \cdot C_i\right)}{\sum_{i=1}^{n} V_i} \qquad \text{(A4.4.1.3)}$$

where

A_n = volume-weighted average of segment n,

V_i = volume of cell i,

C_i = concentration associated with cell i

n = total number of cells

Besides the VWAs, statistical information (mean, variance, standard error) is also generated for the users. VWAs were generated by IDL programs developed in-house. The interpolated grid field data were generated by either distance square inverse or natural-neighbor from samples collected for the LMMBP project.

4.4.1.4 Discussion

It has been observed and documented that extrapolations generated by using the natural-neighbor c-code nngridr could cause problems. Extrapolation sometimes is necessary to estimate values for grid points located outside the convex hull, which is a polygon bounded by the outermost sample data points. At the beginning of the interpolation process of running nngridr, a very large triangle is established which encloses all data being used for interpolation. Then, a pseudo datum is assigned to each of the three vertices of the triangle. Extrapolation, if needed, is performed based on the pseudo data. This process is doomed to be unreliable due to the unpredictable nature of the pseudo numbers and the large triangle used in this process.

This problem can be remedied by adding some extra reasonable pseudo samples at the corners outside the gridding data area so all interpolated grid points will be inside the expanded convex hull. By doing this, nngridr is forced to use interpolation, rather than extrapolation, to calculate estimations based on the original and pseudo samples. This is a more reliable estimation process. The choice of pseudo samples, if necessary, should be based on experience and nearby samples.

Another limitation of nngridr is that it can only handle two-dimensional interpolation. There are occasions when three-dimensional interpolation is needed. One example is the sediment PCB concentration estimations to be used for fish uptake. This is much better represented if the depth of samples could be utilized to define the neighborhood. The neighborhood becomes a three-dimensional ball instead of a two-dimensional circle. Because nngridr is a relatively large program, there was no easy way to add a three-dimensional interpolation.

4.4.1.5 Steps to Run nngridr

The c-code from David Watson, nngridr, was used to generate the natural-neighbor interpolation. Often, nngridr was called within an IDL program to generate the interpolation. Sample data were reformatted to the required IDL format. The interpolation on the 5 km grid was then used for data analysis and visualization (post-process) applications.

Details about how to initialize and run nngridr together with IDL application programs at LLRS follow.

1. Change the c-code

   In nngridr.c, comment out the statement 'Instring()' right after the statement 'printf' ("Change parameters or Make the grid? C or M)", to prevent the read option from a terminal. Therefore, the c-code nngridr to generate the grid is run by using the default option 'M'.

2 Initialization

A. Change the make file and then use the command: %make -f makefile to generate the executable code.

B. The nngridr is run first to generate the initial file and setup parameters (file names, grid configurations, etc.). The result will be in nngridr.ini, which can be used for successive runs without changing parameters again. The most important aspect of initialization is to generate a two-dimensional grid. The southwest corner with longitude -87.9721 and latitude 41.5845 is used as the origin of the grid coordinates. The northeast corner is at longitude -84.7206 and latitude 46.1069 which is the grid coordinate of (53, 102). There is an option to output the grid south-north or north-south. The orientation of output should be set from south-to-north. Otherwise, the image will be upside down. See Watson (1994) for more details.

3. Raw data files are pre-processed to prepare the input data for IDL code. The formats of data files should be the same.

4. The configuration of segmentation should be stored in a file for the segmentation classification.

5. IDL programs are coded to generate data files similar to jdavis.dat by reading the pre-processed data. Coordinates are converted from longitude and latitude coordinates to 5 km grid coordinates.

6. Once the jdavis.dat is established, nngridr is run by the following commands within IDL programs:

   'SPAWN, 'nngridr', Results, /NOSHELL'.

   This creates a child process under the Unix operating system and stores all messages generated by this code into the character array Results.

7 After a successful run (need error checking if run fails), the grid data should be generated and named as nngridr.grd. This file is called in IDL programs to generate contour plots, VWA results, and statistics.

8. For the Unix system, both jdavis.dat and nngridr.grd will be destroyed automatically when new ones are created. For other operating systems, Microsoft Windows, for example, these files need to be deleted.

9. Green Bay data need to be processed separately from open lake data.

References

Research Systems, Incorporated. 1995. IDL User's Guide: Interactive Data Language, Version 4. Research Systems, Incorporated, Boulder, Colorado. 544 pp.

Watson, D. 1994. nngridr - An Implementation of Natural Neighbor Interpolation. Claremont, Australia. 170 pp.